Unsupervised Learning

Letting Data Speak for Itself

Ahmed BADI

Mathematics & Machine Learning Enthusiast

ahmedbadi905@gmail.com
linkedin.com/in/badi-ahmed

December 17, 2025

Overview

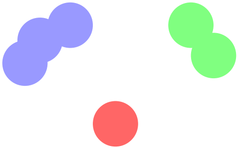
Introduction

What is Unsupervised Learning?

- No labels, no target y.
- Goal: discover hidden structure in $X = \{x_1, \ldots, x_n\}$.
- Typical tasks:
  - Group similar points (clustering).
  - Compress data (dimensionality reduction).
  - Find co-occurrence patterns (association rules).
  - Detect unusual points (anomalies).

The Party Analogy

- You arrive at a party with no name tags.
- After a while, you notice:
    - Group A: talking about football.
    - Group B: discussing cameras.
    - A few people alone (outliers).
- Your brain does clustering, anomaly detection, and compression.



figureClusters and outlier.

Clustering

Clustering Goals

- Partition data into groups (clusters).
- Points in same cluster: similar.
- Points in different clusters: dissimilar.

Main families:

- Centroid-based (k-means).
- Hierarchical (dendrograms).
- Density-based (DBSCAN).

k-means Clustering

Objective

Given K clusters, minimize:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \left\| x_i - \mu_k \right\|^2$$

Algorithm:

1. Initialize K centroids μ_k.
2. Assign each point to nearest centroid.
3. Update each μ_k as mean of its points.
4. Repeat until convergence.

k-means Illustration

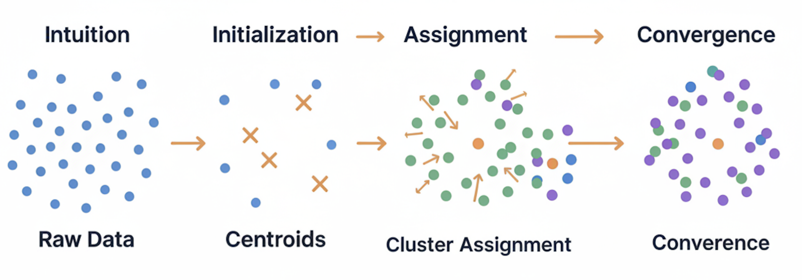


Figure: k-means iterations: assignment and centroid update.

Hierarchical & Density-based Clustering

Hierarchical (Agglomerative)

- Start with each point as its own cluster.
- Merge closest clusters step by step.
- Visualized with a dendrogram.
- Cut at chosen height $\rightarrow K$ clusters.

DBSCAN

- Defines clusters as dense regions.
- Parameters: ε, `minPts`.
- Finds arbitrary shapes.
- Naturally labels noise/outliers.

Dimensionality Reduction

Why Reduce Dimensionality?

- High-dimensional data is hard to visualize and learn from.
- Many features may be redundant or noisy.
- Dimensionality reduction:
  - Compresses data to $q \ll d$.
  - Speeds up models.
  - Helps visualization and understanding.

PCA: Principal Component Analysis

- Finds directions of maximum variance.
- Given covariance matrix Σ:

$$\Sigma v_j = \lambda_j v_j$$

- First components $v_1, \ldots, v_q$ capture most variance.
- Projection: $z_i = V_q^\top x_i$.

PCA Illustration

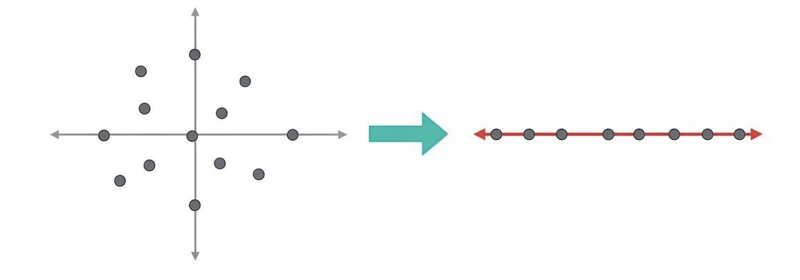


Figure: Data projected onto first principal component.

Nonlinear Methods: t-SNE & UMAP

t-SNE

- Preserves local neighborhoods.
- Great for 2D visualization of clusters.
- Not ideal for global distances or features.

UMAP

- Graph-based manifold learning.
- Often preserves more global structure.
- Fast and scalable for large datasets.

Association Rules

Market Basket Patterns

- Data: transactions T = sets of items.
- Goal: find rules of the form

$$X \Rightarrow Y$$

  e.g. $\{\text{Diapers}\} \Rightarrow \{\text{Beer}\}$.

Quality measures:

- **Support**: frequency of $X \cup Y$.
- **Confidence**: $P(Y|X)$.
- **Lift**: how much X increases chance of Y.

Apriori Algorithm

1. Find all frequent single items (support $\geq$ min_supp).
2. Generate candidate itemsets of size 2, 3, ...
3. Prune candidates whose subsets are not frequent.
4. From frequent itemsets, generate rules with:
   - High confidence.
   - High lift.

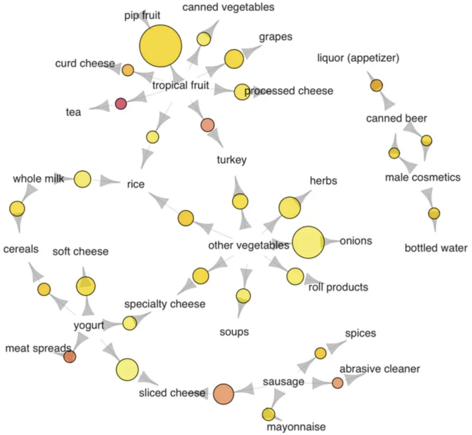


figure From transactions to frequent itemsets and rules.

Beyond the Basics

Anomaly Detection

- Anomalies = rare, unusual points.
- Important for:
  - Fraud detection.
  - Network intrusion.
  - Fault detection.
- Unsupervised approaches:
  - Distance-based (far from neighbors).
  - Density-based (LOF).
  - One-Class SVM, Isolation Forest.

Representation Learning & Autoencoders

- Learn compact latent representation z without labels.
- Autoencoder:

$$x \to \text{Encoder} \to z \to \text{Decoder} \to \hat{x}$$

- If z is low-dimensional:
  - Acts like nonlinear PCA.
  - z can be used for clustering, visualization, or as features.

Conclusion

Advantages & Limitations

Advantages

- No labels needed.
- Strong exploratory power.
- Useful preprocessing for supervised ML.

Limitations

- No ground truth for evaluation.
- Sensitive to hyperparameters.
- Results may be hard to interpret.

Thank You!

"Let the data tell you its story."

Questions?